



Live blog summarization

P. V. S. Avinesh¹  · Maxime Peyrard¹ ·
Christian M. Meyer¹ 

Accepted: 13 October 2020 / Published online: 2 January 2021
© The Author(s) 2021

Abstract Live blogs are an increasingly popular news format to cover breaking news and live events in online journalism. Online news websites around the world are using this medium to give their readers a minute by minute update on an event. Good summaries enhance the value of the live blogs for a reader, but are often not available. In this article, (a) we first define the task of summarizing a live blog, (b) study ways of automatically collecting corpora for live blog summarization, and (c) understand the complexity of the task by empirically evaluating well-known state-of-the-art unsupervised and supervised summarization systems on our new corpus. We show that live blog summarization poses new challenges in the field of news summarization, since frequency and positional signals cannot be used. We make our tools publicly available to reconstruct the corpus and to conduct our empirical experiments. This encourages the research community to build upon and replicate our results.

Keywords Live blog summarization · Corpus construction · Focused crawling · Online journalism

✉ P. V. S. Avinesh
avinesh.pvs@gmail.com;
https://www.aiphes.tu-darmstadt.de

Maxime Peyrard
maxime.peyrard@epfl.ch;
https://www.aiphes.tu-darmstadt.de

Christian M. Meyer
chmeyer.de@gmail.com;
https://www.aiphes.tu-darmstadt.de

¹ Research Training Group AIPHES and UKP Lab, Computer Science Department, Technische Universität Darmstadt, Darmstadt, Germany

1 Introduction

Live blogs are dynamic news articles providing a rolling textual coverage of an ongoing event. One or multiple journalists continually post micro-updates about the event, which are displayed in chronological order. The updates contain a wide variety of modalities and genres, including text, video, audio, images, social media excerpts, and external links. During the last 5 years, live-blogging emerged as a very popular way to disseminate news offered by many major news organizations, such as the *BBC*, *The Guardian*, or *The New York Times*.

Several different kinds of events are regularly covered by live blogs, including sport games, elections, ceremonies, protests, conflicts, and natural disasters. Thurman and Schapals (2017, p. 1) report a journalist's view that "live blogs have transformed the way we think about news, our sourcing, and everything". Besides their timeliness, live blogs differ from common news articles by utilizing more original sources and providing information as smaller chunks, often written in a different tone than in traditional news writing (Thurman and Walters 2013).

Figure 1 shows an example of a live blog on the constitution of a new Brexit committee provided by *The Guardian*.¹ Live blogs typically consist of metadata, such as date, title, and authors and a list of postings with the updated information. For larger events, journalists provide intermediate summaries shown at the top of the article. At the end of the broadcasting, a journalist usually aggregates the postings and, if available, intermediate summaries to present the most important information about the event as timelines, short texts, or bullet point lists to the users. Figure 2 shows an excerpt of a completed live blog by the *BBC* which consists of 360 postings (distributed over 19 pages) and a summary shown as four bullet point items.²

In this work, we propose to leverage these human-written summaries to investigate the novel task of automatic live blog summarization. To this end, we provide a new corpus construction approach for producing a dataset of live blogs, and we evaluate state-of-the-art summarization systems for this new summarization task. Our work has multiple direct applications in digital journalism and news research, since automatic summarization tools for live blogs help journalists to save time during live-blogging and enable instant updates of the intermediate summaries on a live event. However, the automatic live blog summarization task also comes with new challenges:

1. Unlike a news article, the postings of a live blog do not form one coherent piece of text. Instead, each posting introduces facts or opinions from a single source which might be highly or only marginally related to the overarching topic. For example, the live blog in Fig. 2 contains a posting commenting the relationship between Theresa May and Angela Merkel, which is related to the overall Brexit topic, but not to the Supreme Court case. In similar lines, the live blog contains

¹ <https://www.theguardian.com/politics/blog/live/2019/jan/07/brexit-latest-commons-vote-boris-johnson-claims-no-deal-is-closest-to-what-people-voted-for-politics-live> (accessed January 7, 2019).

² <https://www.bbc.com/news/live/uk-politics-37976580/> (accessed January 7, 2019).

Politics live with Andrew Sparrow
Sparrow Politics

May to chair new cabinet committee on Brexit planning, including for no deal - Politics live

● LIVE Updated 5m ago

Rolling coverage of the day's political developments as they happen

- No-deal Brexit rehearsal tests traffic congestion in Kent
- Germany and Ireland step up efforts to find Brexit border 'fix'
- May will win vote on her deal on 15 January, says Brexit minister


Andrew Sparrow
@AndrewSparrow
Mon 7 Jan 2019 13:42 GMT

96 3,771

2h ago
Theresa May's speech and Q&A


2h ago
May to chair new cabinet committee on Brexit planning, including for no deal

4h ago
Boris Johnson claims no-deal Brexit is 'closest to what people voted for'



▲ Theresa May (3rd from left), health secretary Matt Hancock (left) and NHS England chief executive Simon Stevens (centre) visiting the wards at Alder Hey Children's Hospital, Liverpool this morning. Photograph: Charlotte Graham/Daily Telegraph/PA

3m ago
13:42



▲ Former Tory minister and Hong Kong governor Lord Patten speaking during a People's Vote event at Coin Street Neighbourhood Centre, central London. Photograph: Kinsty O'Connor/PA

10m ago
13:37

Two of the main anti-Brexit groups have put out press notices claiming there was something inappropriate about **Theresa May** defending her Brexit plans on a visit to Alder Hey children's hospital in Liverpool this morning.

Best for Britain says 13% of the hospital's doctors are EU nationals (or non-British EU nationals, to be precise). It has released this comment from the Labour MP **Alison McGovern**.

▲▲ *As May parades her NHS 10-year plan at Alder Hey hospital, our health service is facing the greatest threat to its existence. World-class children's hospitals like Alder Hey are held together by the dedication and expertise of EU staff, who we cannot afford to lose due to **Brexit**.*

And the **People's Vote** campaign says the hospital was built with £56m in funding from the European Investment Bank. It released this comment from the Labour MP **Luciana Berger**.

▲▲ *It is beyond parody that the prime minister has the audacity to claim that Brexit benefits our NHS [see [12.13pm](#)], standing in a hospital that was built using over £50m of financing available to the UK because of our EU membership.*

Access to this funding is vital. NHS trusts across the country rely on European investment in order to build the health facilities we need. The government willingly cutting off access to this - especially with absolutely no plan for how to replicate it - amounts to a dereliction of duty.

This is further proof that Brexit means less money for our NHS, not more. The fibs people were told during the referendum in 2016 are proven wrong every day. This is why we need a People's Vote.

Title and domain

Date, author, and space for intermediate summary

Posting 1

Posting 2

Fig. 1 Live blog example from *The Guardian* (two newest postings visible)

Last day of Supreme Court Brexit case

5 Dec 2016 06:00

Read more: [Court 'won't overturn Brexit vote'](#) - [All you need to know about Brexit](#)

Summary

- Supreme Court case ends with reminder it's not about stopping Brexit
- Government appealed against ruling it needs MPs' approval to trigger Brexit
- Judgement is expected in January
- Watch highlights of each day via clips above, or scroll down to see how events unfolded

Live Reporting

By Jackie Storer and Alex Hunt


23:23 8 Dec 2016

Watch: Highlights of Thursday at Supreme Court

[f](#)
[t](#)
[Share](#)

18:30 8 Dec 2016

Supreme Court 'won't overturn Brexit'



The Brexit hearing draws to a close with a reminder the court will not overturn the referendum.


[Read more >](#)

[f](#)
[t](#)
[Share](#)

ADVERTISEMENT

18:10 8 Dec 2016

'A fine example of the rule of law and British constitution in action'



Clive Coleman

BBC legal correspondent

This was a case about a hugely important point of pure constitutional law, but it arrived at the Supreme court in a blizzard of politics, acrimony, threats of violence against Gina Miller the woman at its heart, and some very personal press criticism of the judges about to hear it.

Like many Supreme Court cases it rapidly took on the feel of an academic seminar, often inaccessible and at times impenetrable to the non-lawyer.

Not a ratings winner then, but it has been a fine example of the rule of law and the British constitution in action. Independent judges considering issues raised by citizens, exercising their right to have a decision of ministers scrutinised by a court to determine whether it is lawful or not.

The proceedings have been unfailingly courteous and all of the key players have had their arguments carefully considered. That is not something that happens in every country around the world. So, even if the nation remains divided on the issue, it can take pride in the process by which it is being determined.

[f](#)
[t](#)
[Share](#)

Summary
Postings

Fig. 2 Archived live blog example from the *BBC* (three newest postings visible)

- multiple topic shifts (e.g., focusing on the MP's opinions or the government appeal). This lets us assume that single-document summarizers cannot be used out of the box.
2. A particular challenge is that positional features cannot be used to estimate information importance, because live blogs are chronologically ordered and, unlike news articles, do not necessarily report the most important information first. Thus, baselines that extract the first few sentences or single-document

summarization approaches building extensively on the position of a sentence are not suitable for live blog summarization.

3. The postings of live blogs are very heterogeneous, covering multiple genres, modalities, and styles. They also differ in their length and, unlike most multi-document summarization datasets, they are hardly redundant. Automatic live blog summarization approaches therefore have to deal with heterogeneous data and identify novel ways of judging importance that are not solely based on the frequency signal.

In summary, live blog summarization is a special kind of multi-document summarization, but faces highly heterogeneous, temporally ordered input. It is similar to update summarization, but has to deal with low redundancy and occasional topic shifts. Moreover, it is related to real-time summarization, where summaries are to be created without having full information about the topic yet.

The remaining article is structured as follows: Sect. 2 discusses related work on live blog summarization and summarization corpora. In Sect. 3, we introduce our first contribution by suggesting a novel pipeline to collect and extract the human-written summaries and postings from online live blogs, which we make available as open-source software from our GitHub repository.³ Section 4 provides a detailed analysis of the corpus we created from live blogs of two major news publishers, the *BBC* and *The Guardian*, using our pipeline. Sections 5 and 6 describe our second contribution, as we propose the new task of live blog summarization and benchmark our corpus with multiple commonly used summarization methods. While we find that live blog summarization is a challenging task, our work aims at stipulating further research in this area, for which we provide both reference data and benchmark results. Our results show that off-the-shelf summarization systems are not effective for live blog summarization, as they do not properly take into account the large number of heterogeneous postings of a live blog. Section 7 concludes our work and points to multiple directions for future research.

2 Related work

In this section, we discuss previous work on summarization corpora and automatic summarization methods as well as journalistic NLP applications related to the task of live blog summarization.

2.1 Summarization corpora

The most widely used summarization corpora have been published in the Document Understanding Conference⁴ (DUC) series. In total, there are 139 document clusters with 376 human-written reference summaries across DUC 2001, 2002, and 2004. Although the research community has often used these corpora, their limited size

³ <https://github.com/AIPHES/live-blog-summarization>.

⁴ <http://duc.nist.gov/>.

prevents training advanced methods, such as encoder–decoder architectures, and it is time-consuming and labor-intensive to extend such corpora with large numbers of manually written summaries.

Large datasets exist particularly for single-document summarization tasks, including the ACL Anthology Reference Corpus (Bird et al. 2008) and the CNN/Daily Mail dataset (Hermann et al. 2015). The latter contains large pairs of 312k online news articles and multi-sentence summaries used for neural summarization approaches (Nallapati et al. 2016; See et al. 2017). However, this dataset contains only one source document, whereas live blogs have a larger number of postings (typically more than 100) that act like individual small documents.

Another recent work uses social media posts on Twitter to create large-scale multi-document summaries for news: Cao et al. (2016) use hashtags to cluster the tweets on the same topic, and they assume the tweet's content to be a reference summary for the document linked by the tweet. Their corpus consists of 204 document clusters with 1114 documents and 4658 reference tweets. Lloret and Palomar (2013) create a similar corpus of English and Spanish news documents and corresponding tweets linking to them.

Other multi-document summarization datasets focus on heterogeneous sources: Zopf et al. (2016) and Zopf (2018) use Wikipedia articles as reference summaries and automatically search for potential source documents on the web. Benikova et al. (2016) propose an expert-based annotation setup for creating a summarization corpus for highly heterogeneous text genres from the educational domain. In similar lines of research, Tauchmann et al. (2018) use a combination of crowdsourcing and expert annotation to create a hierarchical summaries for a heterogeneous web crawl. Giannakopoulos et al. (2015) discuss multilingual summarization corpora and Li et al. (2017) introduce a corpus of reader-aware multi-document summaries, which jointly aggregate news documents and reader comments.

2.2 Automatic summarization

2.2.1 Extractive summarization

Until recently (Yao et al. 2017), the vast majority of research focused on extractive summarization, which outputs a selection of important sentences or phrases available in the input sources (Ko and Seo 2008; Nenkova and McKeown 2012). By selecting already grammatical elements, extractive summarization reduces to a combinatorial optimization problem (McDonald 2007). To solve such combinatorial problems, summarization systems have leveraged powerful techniques like Integer Linear Programming (ILP) or submodular maximization.

In order to score sentences and phrases, Luhn (1958) initially introduced the simple, but influential idea that sentences containing the most important words are most likely to embody the original document. This hypothesis was experimentally supported by Nenkova et al. (2006), who showed that humans tend to use words appearing frequently in the sources to produce their summaries. Many subsequent works exploited and refined this strategy. For instance, by computing TF·IDF (Spark Jones 1972) or likelihood ratio (Dunning 1993).

Words serve as a proxy to represent the topics discussed in the sources. However, different words with a similar meaning may refer to the same topic and should not be counted separately. This observation gave rise to a set of important techniques based on topic models (Allaahyari et al. 2017). These approaches can be divided into sentence clustering (Radev et al. 2000), Latent Semantic Analysis (Deerwester et al. 1990; Gong and Liu 2001), and Bayesian topic models (Blei et al. 2003).

Graph-based methods form another powerful class of approaches which combine repetitions at the word and at the sentence level. They were developed to estimate sentence importance based on word and sentence similarities (Mani and Bloedorn 1997, 1999; Mihalcea and Tarau 2004). One of the most prominent examples is LexRank (Erkan and Radev 2004), which we run on our dataset in Sect. 6.

More generally, many indicators for sentence importance were proposed and therefore the idea of combining them to develop stronger indicators emerged (Aone et al. 1995). Kupiec et al. (1995) suggested that statistical analysis of summarization corpora would reveal the best combination of features. For example, the frequency computation of words or n-grams can be replaced with learned weights (Hong and Nenkova 2014; Li et al. 2013). Additionally, structured output learning permits to score smaller units while providing supervision at the summary level (Li et al. 2009; Peyrard and Eckle-Kohler 2017).

A variety of works proposed to learn importance scores for sentences (Yin and Pei 2015; Cao et al. 2015). This started a huge body of research comparing different learning algorithms, features and training data (Hakkani-Tur and Tur 2007; Hovy and Lin 1999; Wong et al. 2008). Nowadays, sequence-to-sequence methods are usually employed (Nallapati et al. 2017; Kedzie et al. 2018). These approaches are presented in Sect. 5 and tested on live blog summarization in Sect. 6.

2.2.2 *Abstractive summarization*

In contrast to extractive summarization, abstractive summarization aims to produce new and original texts (Khan et al. 2016) either from scratch (Rush et al. 2015; Chopra et al. 2016), by fusion of extracted parts (Barzilay and McKeown 2005; Filippova 2010), or by combining and compressing sentences from the input documents (Knight and Marcu 2000; Radev et al. 2002). Intuitively, abstractive systems have more degrees of freedom. Indeed, careful word choices, reformulation and generalization should allow condensing more information in the final summary.

Recently, end-to-end training based on the encoder-decoder framework with long short-term memory (LSTM) has achieved huge success in sequence transduction tasks like machine translation (Sutskever et al. 2014). For abstractive summarization, large single-document summarization datasets rendered possible the application of such techniques. For instance, (Rush et al. 2015) introduced a sequence-to-sequence model for sentence simplification. Later, Chopra et al. (2016) and Nallapati et al. (2016) extended this work with attention mechanisms. Since words from the summary are often retained from the original source, copy mechanisms (Gu et al. 2016; Gulcehre et al. 2016) have been thoroughly investigated (Nallapati et al. 2016; See et al. 2017).

2.2.3 Update summarization

After the DUC series, the Text Analysis Conference⁵ (TAC) series introduced the update summarization task (Dang and Owczarzak 2008). In this task, two summaries are provided for two sets of documents and the summary of the second set of documents is an update of the first set. Although the importance of text to be included in the summary solely depends on the novelty of the information, the task usually observes only a single topic shift. In live blogs, however, there are multiple sub-topics and the importance of the sub-topics changes over time.

2.2.4 Real-time summarization

Real-time summarization began at the Text REtrieval Conference⁶ (TREC) 2016 and represents an amalgam of the microblog track and the temporal summarization track (Lin et al. 2016). In real-time summarization, the goal is to automatically monitor the stream of documents to keep a user up to date on topics of interest and create email digests that summarize the events of that day for their interest profile. The drawback of this task is that they have a predefined time frame for evaluation due to the real-time constraint, which makes the development of systems and replicating results arduous. Note that live blog summarization is very similar to real-time summarization, as the real-time constraint also holds true for live blogs if the summarization system is applied to the stream of postings. Moreover, the Guardian live blogs do consist of updated and real-time summaries, but this requires different real-time crawling strategies which are out of the scope of this work.

2.2.5 Multi-tweet summarization

Tweets are 140-character short messages shared on Twitter, a micro-blogging website with a large number of users contributing and sharing content. Multi-tweet summarization allows the users to quickly grasp the gist of the large number of tweets. For multi-tweet summarization, previous work employed graph-based approaches (Liu et al. 2012) similar to LexRank, Hybrid TF-IDF (Sharifi et al. 2010) which ranks tweets based on TF-IDF, and ILP (Cao et al. 2016; Liu et al. 2011) optimizing the coverage of information in the summary. Summarizing tweets is similar to live blog summarization, since the postings of live blogs are similarly structured as tweets, but typically use more formal language than on Twitter. The postings of live blogs are also more heterogeneous, as tweets can be part of a live blog along with many other types of postings, such as images, interviews, or reporting. In our work, we benchmark our live blog summarization corpus with similar approaches, including graph-based, TF-IDF, and ILP-based methods.

⁵ <http://www.nist.gov/tac/>.

⁶ <http://trec.nist.gov/>.

2.3 NLP and journalism

Leveraging natural language processing methods for journalism is an emerging research topic. The SciCAR conferences⁷ and the recent “Natural Language Processing meets Journalism” workshops (Birnbaum et al. 2016; Popescu and Strapparava 2017, 2018) are predominant examples for this development. Previous research focuses on news headline generation and click-bait analysis (Blom and Hansen 2015; Gatti et al. 2016; Szymanski et al. 2016), abusive language and comment moderation (Clarke and Grieve 2017; Kolhatkar and Taboada 2017; Pavlopoulos et al. 2017; Schmidt and Wiegand 2017), news bias and filter bubble analyses (Baumer et al. 2015; Bozdag and van den Hoven 2015; Fu et al. 2016; Kuang and Davison 2016; Potash et al. 2017), as well as news verification and fake news detection (Brandtzaeg et al. 2015; Thorne et al. 2017; Bourgonje et al. 2017; Hanselowski et al. 2018; Thorne et al. 2018). We are not aware of any work on live blog summarization or computational approaches closely related to journalistic live blogging.

Live blogs as such have been previously discussed in the domain of digital journalism. Thorsen (2013) gives a general introduction about challenges and opportunities of live blogging. Thurman and Walters (2013) and Thurman and Newman (2014) study the production processes and the readers’ consumption behavior, Thurman and Schapals (2017) evaluate aspects of transparency and objectivity, and Thorsen and Jackson (2018) analyze sourcing practices in live blogs. Further works discuss certain types of live blogs, such as live blogs on sport events (McEnnis 2016) or terrorist attacks (Wilczek and Blangetti 2018). None of these works focuses on intermediate or final summaries in live blogs or computational approaches to assist the journalists.

3 Corpus construction pipeline

In this section, we describe the three steps to construct our live blogs summarization corpus: (1) live blog crawling yielding a list of URLs, (2) content parsing and processing, where the documents and corresponding summaries with the metadata are extracted from the URLs and stored in a JSON format, and (3) live blog pruning as a final step for creating a high-quality gold standard live blog summarization corpus.

3.1 Live blog crawling

A frequently updated index webpage⁸ references all archived live blogs of the Guardian. We take a snapshot of this page yielding 16,246 unique live blog URLs. In contrast, the BBC website has no such live blog archive. Thus, we use an iterative approach similar to BootCaT (Baroni and Bernardini 2004) to bootstrap our corpus.

⁷ <https://www.scicar.de>.

⁸ <http://www.theguardian.com/tone/minutebyminute>.

Table 1 Initial BBC live blogs links used to extract seed terms

Title	URL
Politics round-up: 6 July	http://www.bbc.com/news/live/uk-politics-33406777
Over £36bn wiped off FTSE	https://www.bbc.com/news/live/business-34358976
Stormont	https://www.bbc.com/news/live/uk-northern-ireland-politics-35640347
Africa highlights	http://www.bbc.com/news/live/world-africa-35518162
Election live—7 April	https://www.bbc.com/news/live/election-2015-32170452
School report practice	http://www.bbc.com/news/live/education-31313670
IPCC report launch	https://www.bbc.com/news/live/science-environment-29820051
Junior doctor's strike	http://www.bbc.com/news/live/health-35290222
Search for Flight QZ8501	http://www.bbc.com/news/live/world-asia-30630322
Oregon shooting	http://www.bbc.com/news/live/world-us-canada-34420055

Algorithm 1 shows pseudo code for our iterative crawling approach, which is based on a small set of live blog URLs L_0 shown in Table 1. From these live blogs, we extract a set of seed terms K_0 using the 500 terms with the highest TF-IDF scores. Table 2 shows K_0 for our corpus. The iterative procedure uses the seed terms K_0 to gather new live blog URLs by issuing automated Bing queries⁹ created using recurring URL patterns P for live blogs (line 7). We collect all valid links returned by the Bing search (line 8) and extract new key terms K_t from each crawled live blog (line 12). Similar to the seed terms, we define K_t as the top 500 terms sorted by TF-IDF. The new key terms are then used to generate the Bing queries in the subsequent iterations (line 7). The process is repeated until no new live blogs are discovered anymore (line 9). For our corpus, we use the pattern

site : http : //www.bbc.com/news/live/ <keyterm >

where <key term> is one of the extracted key terms K_{t-1} from the previous iteration (or the seed terms if $t = 1$).

Using the proposed algorithm, we run 4000 search queries returning each around 1000 results on average, from which we collected 9931 unique URLs. Although our method collects a majority of the live blogs in the 4000 search queries, a more sophisticated key terms selection could minimize the search queries and maximize the unique URLs. An important point to note is that we find the collected BBC live blog URLs predominantly cover more recent years. This usage could be due to the Bing Search API preferring recent articles for the first 100 results.

By choosing a different set of seed URLs L_0 or seed terms K_0 and different URL patterns P , our methodology can be applied to other news websites featuring live blogs, such as *The New York Times*, the *Washington Post* or the German *Spiegel*.

⁹ <https://azure.microsoft.com/en-us/services/cognitive-services/bing-web-search-api>.

3.2 Content parsing and processing

Once the URLs are retrieved, we fetch the HTML content, remove the boiler-plate using the BeautifulSoup¹⁰ parser and store the cleaned data in a JSON file. During this step, unreachable URLs were filtered out. We discard live blogs for which we could not retrieve the summary or correctly parse the postings.

We parse metadata, such as URL, author, date, genre, summaries, and all postings for each live blog using site-specific regular expressions on the HTML source files. The automatic extraction is generally difficult, as the markup structure may change over time. For BBC live blogs, both the postings and the bullet-point summaries follow a consistent pattern, we can easily extract automatically. For the Guardian, we identify several recurring patterns which cover most of the live blogs. The Guardian provides live blogs since 2001, but they were in an experimental phase until 2008. Due to the lack of a specific structure or a summary during this experimental phase,

Algorithm 1 Iterative Live blog crawling

```

1: input Seed URLs  $L_0$ , URL patterns  $P$ 
2: output List of live blog URLs  $L$ 
3: procedure CRAWLLIVEBLOGS
4:    $L \leftarrow L_0$ 
5:    $K_0 \leftarrow \text{extractKeyTerms}(L_0)$ 
6:   for  $t = 1 \dots T$  do
7:      $Q_t \leftarrow \text{createQueries}(K_{t-1}, P)$ 
8:      $L_t \leftarrow \text{obtainLinks}(Q_t)$ 
9:     if  $L \cup L_t = L$  then
10:      return  $L$ 
11:   else
12:      $K_t \leftarrow \text{extractKeyTerms}(L_t) - \bigcup_{i=0}^{t-1} K_i$ 
13:      $L \leftarrow L \cup L_t$ 
14:   end if
15: end for
16: return  $L$ 
17: end procedure

```

we had to remove about 10k of the crawled live blogs, for which we could not automatically identify the postings or the summary. However, after 2008, the live blogs showed a consistent structure, as they received a prominent place in the web site. After this step, 7307 live blogs remain for the BBC and 6450 for the Guardian.

3.3 Live blog pruning

To further clean the data, we remove live blogs covering multiple topics, as they can be quite noisy. For example, BBC provides some live blogs discussing all events

¹⁰ <https://pypi.org/project/beautifulsoup4/>.

Table 2 Sample seed terms extracted from the initial ten BBC live blogs

World	Technology	UK	Business	Politics	Health
Education	Science	Environment	Africa	Asia	Europe
Latin America	Middle East	US and Canada	Northern Ireland	Scotland	NHS
Nottingham	Headlines	Issues	Justice	Royal	Crime
Rangers	Details	Risk	Emergency	Food	Bid
Essex	Traffic	Updates	Oxford	Schools	Commons
Officer	Birmingham	Amendment	National	Investment	Investigation
Safety	Sheffield	Appeal	Jobs	Northampton	Residents
Workers	Scene	Community	Midlands	Authority	Spending
Evidence	Law	Housing	Concerns	Impact	Charges

happening in a certain region within a given time frame (e.g., *Essex: Latest updates*). We also prune live blogs about sport games and live chats, because their summaries are based on simple, easy-to-replicate templates.

We further prune live blogs based on their summaries. We first remove a sentence of a summary if it has less than three words. Then, we discard live blogs whose summaries have less than three sentences. This is to ensure the quality of the corpus, since overly short summaries would yield a different summarization goal similar to headline generation and they are typically an indicator for a non-standard live blog layout in which the summary has been separated to multiple parts of the website.

After the whole pruning step, 762 live blogs remained for BBC and 1683 for the Guardian. Overall, 10% of the initial set of live blogs, both for BBC and the Guardian remain after our selective pruning. This is to ensure high-quality summaries for the live blogs. Although the pruning rejects 90% of the live blogs, the size of the live blog corpus is still 20–30 times larger than the classical corpora released during DUC, TREC, and TAC tasks (Table 3).

3.4 Code repository

We publish our tools for reconstructing the live blog corpus as open-source software under the Apache License 2.0 on GitHub.¹¹ This repository helps to replicate our results and advance research in live blog summarization.

The repository consists of (a) raw and pruned URL lists, (b) tools for crawling live blogs, (c) tools for parsing the content of the URLs and transforming the results into JSON, and (d) code for computing benchmark results and corpus statistics.

¹¹ <https://github.com/AIPHES/live-blog-summarization>.

Table 3 Number of live blogs for BBC and the Guardian after each step of our pipeline

Source	Crawling	Processing	Pruning
BBC	9931	7307	762
Guardian	16,246	6405	1683
Total corpus	26,177	13,712	2655

4 Corpus analysis

Our final corpus yields a multi-document summarization corpus, in which the individual topics correspond to the crawled live blogs and the set of documents per topic corresponds to the postings of the live blog. We compute several statistics about our corpus and report them in Table 4. The number of postings per live blog is around 95 for BBC and 56 for the Guardian. In comparison, standard multi-document summarization datasets like DUC 2004¹² and TAC 2008A¹³ have only 10 documents per topic. Furthermore, we observe that the postings are quite short as there is an average of 62 words per posting for BBC and 108 for the Guardian. The summaries are also shorter than the summaries of standard datasets: The summaries of DUC 2004 and TAC 2008A are expected to contain 100 words. However, our final corpus is larger overall, because it contains 2655 live blogs (i.e., topics) and 186,999 postings (i.e., documents). With that many data points, machine learning approaches become readily applicable.

4.1 Domain distribution

The live blogs in our corpus cover a wide range of subjects from multiple domains. In Table 5, we report the distribution across all domains in the final corpus (BBC and Guardian combined). While we observe that politics, business, and news are the most prominent domains, there is also a number of well-represented domains, such as local and international events or culture.

4.2 Heterogeneity

The resulting corpus is expected of exhibiting various levels of heterogeneity. Indeed, it contains live blogs with mixed writing styles (short and to the point vs. longer descriptive postings, informal language, quotations, encyclopedic background information, opinionated discussions, etc.). Furthermore, live blogs are subject to topic shifts which can be observed by changes in words usage.

To measure this textual heterogeneity, we use information theoretic metrics on word probability distributions like it was done before in analyzing the heterogeneity of summarization corpora (Zopf et al. 2016). Based on the Jensen-Shannon (JS)

¹² <http://duc.nist.gov/duc2004>.

¹³ <https://tac.nist.gov/2008>.

Table 4 Corpus statistics for BBC and the Guardian live blogs

Statistic	BBC	Guardian
Number of live blogs	762	1683
Number of postings	92,537	94,462
Average postings per live blog	95.01	56.19
Average words per posting	61.75	107.53
Average words per summary	59.48	42.23

Table 5 Domain distribution of our final corpus

Domain	Live blogs	Proportion (%)
Politics	834	31.41
Business	421	15.86
General news	369	13.90
UK local events	368	13.86
International events	337	12.69
Culture	186	7.01
Science	60	2.26
Society	27	1.02
Others	53	2.00

Table 6 Average textual heterogeneity of our corpora compared to standard datasets

	BBC	Guardian	DUC 2004	TAC 2008A
TH_{JS}	0.5917	0.5689	0.3019	0.3188

divergence, they defined a measure of textual heterogeneity TH for a topic T composed of documents d_1, \dots, d_n as

$$TH_{JS}(T) = \frac{1}{n} \sum_{d_i \in T} JS(P_{d_i}, P_{T \setminus d_i}) \quad (1)$$

Here P_{d_i} is the frequency distribution of words in document d_i and $P_{T \setminus d_i}$ is the frequency distribution of words in all other documents of the topic except d_i . The final quantity TH_{JS} is the average divergence of documents with all the others and provides, therefore, a measure of diversity among documents of a given topic.

We report the results in Table 6. To put the numbers in perspective, we also report the textual heterogeneity of the two standard multi-document summarization corpora DUC 2004 and TAC 2008A. The heterogeneity in BBC and Guardian are similar. Thus, heterogeneity of our corpus is much higher than in DUC 2004 and TAC 2008A, indicating that our corpus contains more lexical variation inside its topics.

4.3 Compression ratio

Additional factors which determine the difficulty of the summarization task are the length of the source documents and the summary (Nenkova and Louis 2008). The input document sizes of the BBC and the Guardian are on an average 5890 and 6048 words, whereas the summary sizes are only around 59 and 42 words respectively. In contrast, typical multi-document DUC datasets have a much lower compression ratio, since their input documents have on average only 700 words, while the summaries have 100 words. Thus, we expect that the high compression ratio makes live blog summarization even more challenging.

5 Automatic summarization methods

To automatically summarize live blogs, we employ methods that have been successfully used for both single and multi-document summarization. Some variants of them have also been applied to update summarization tasks.

5.1 Unsupervised methods

5.1.1 *TF-IDF*

Luhn (1958) scores sentences with the term frequency and the inverse document frequency (TF-IDF) of the words they contain. The best sentences are then greedily extracted.

5.1.2 *LexRank*

Erkan and Radev (2004) constructs a similarity graph $G(V, E)$ with the set of sentences V and edges $e_{ij} \in E$ between two sentences v_i and v_j if and only if the cosine similarity between them is above a given threshold. Sentences are then scored according to their PageRank in G .

5.1.3 *LSA*

Steinberger and Jezek (2004) computes a dimensionality reduction of the term-document matrix via singular value decomposition (SVD). The sentences extracted should cover the most important latent topics.

5.1.4 *KL-Greedy*

Haghighi and Vanderwende (2009) minimizes the Kullback-Leibler (KL) divergence between the word distributions of the summary and the documents.

5.1.5 ICSI

Gillick and Favre (2009) propose using global linear optimization to extract a summary by solving a maximum coverage problem considering the most frequent bigrams in the source documents. ICSI has been among the state-of-the-art MDS systems when evaluated with ROUGE (Hong et al. 2014).

ICSI's concept-based summarization can be formalized using an Integer Linear Programming (ILP) framework. Let C be the set of concepts in a given set of source documents D , c_i the presence of the concept i in the resulting summary, w_i a concept's weight, ℓ_j the length of sentence j , s_j the presence of sentence j in the summary, and Occ_{ij} the occurrence of concept i in sentence j . Based on these definitions, the following ILP has to be solved:

$$\text{Maximize } \sum_i w_i c_i \quad (2)$$

$$\text{subject to } \forall j. \sum_j \ell_j s_j \leq L \quad (3)$$

$$\forall i, j. s_j Occ_{ij} \leq c_i \quad (4)$$

$$\forall i. \sum_j s_j Occ_{ij} \geq c_i \quad (5)$$

$$\forall i. c_i \in \{0, 1\} \quad (6)$$

$$\forall j. s_j \in \{0, 1\} \quad (7)$$

The objective function (2) maximizes the occurrence of concepts c_i (typically bigrams) in the summary based on their weights w_i (e.g., document frequency). The constraint formalized in (3) ensures that the summary length is restricted to a maximum length L , (4) ensures the selection of all concepts in a sentence s_j if s_j has been selected for the summary. Constraint (5) ensures that a concept is only selected if it is present in at least one of the selected sentences.

5.2 Supervised methods

The supervised extractive summarization task as a sequence labeling problem using the formulation by Conroy and O'Leary (2001): Given a document set containing n sentences $(s_1, \dots, s_i, \dots, s_n)$, the goal is to generate a summary by predicting a label sequence $(y_1, \dots, y_i, \dots, y_n) \in \{0, 1\}^n$ corresponding to the n sentences, where $y_i = 1$ indicates that the i -th sentence is included in the summary. The summaries are constructed with a word budget L , which enforces a constraint on the summary length $\sum_{i=1}^n y_i \cdot |s_i| \leq L$. Figure 3 shows the neural network architecture of the four state-of-the-art sentence extractors we describe below.

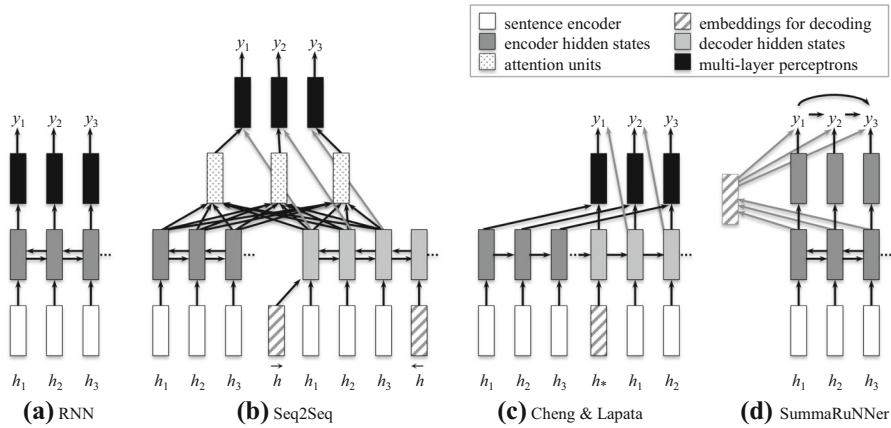


Fig. 3 Architectures of the sentence extractors RNN, Seq2Seq, Cheng and Lapata, and SummaRuNNer

5.2.1 RNN

Kedzie et al. (2018) propose a simple bidirectional RNN-based tagging model. In the sentence encoder, the forward and backward outputs of each sentence are passed through a multi-layer perceptron with sigmoid function as the output layer to predict the probability of extracting each sentence.

5.2.2 Seq2Seq

In the same paper, Kedzie et al. (2018) also propose a sequence-to-sequence (Seq2Seq) extractor which tackles the shortcoming of the RNN extractor i.e. the inability to capture long range dependencies between the sentences. The Seq2Seq extractor thus uses an attention mechanism (Bahdanau et al. 2015; See et al. 2017; Rush et al. 2015) popularly used in machine translation and abstractive summarization. The Seq2Seq extractor is divided into encoder and decoder, where the sentence embeddings are first encoded by a bidirectional GRU and a separate decoder GRU transforms each sentence into a query vector. The query vector attends to the encoder output and is concatenated with the decoder GRU’s output. These concatenated outputs are then fed into a multi-layer perceptron to compute the probabilities for extraction.

5.2.3 Cheng and Lapata

Cheng and Lapata (2016) propose a Seq2Seq model where the encoder RNN is fed with the sentence embedding and the final encoder state is passed on to the first step of the decoder RNN. The decoder takes the same sentence embeddings as input and the outputs are used to predict the y_i labels defining the summary. To induce dependencies of y_i on $y_{<i}$, the decoder input is weighted by the previous extraction probabilities $y_{<i}$.

5.2.4 SummaRuNNer

Nallapati et al. (2017) propose a sentence extractor where the sentence embeddings are passed into a bidirectional RNN and the output is concatenated. Then, they average the RNN output to construct a document representation, and they sum up the previous RNN outputs weighted by extraction probabilities to construct a summary representation for each time step. Finally, the extraction probabilities are calculated using the document representation, the sentence position, the RNN outputs, and the summary representation at the i -th step. The iterative summary representation process intuitively considers dependencies of y_i on all $y_{<i}$.

We test each sentence extractor with two input encoders that compute sentence representations based on the sequence of word embeddings.

5.2.5 Averaging encoder (Avg)

The averaging encoder creates sentence representations

$$h_i = \frac{1}{|s_i|} \sum_{j=1}^{|s_i|} w_j$$

by averaging the word embeddings $(w_1, \dots, w_j, \dots, w_{|s_i|})$ of a sentence s_i .

5.2.6 CNN encoder

The CNN sentence encoder employs a series of one-dimensional convolutions over word embeddings, which is similar to the architecture proposed by Kim (2014) used for text classification. The final sentence representation h_i is the concatenation of the max-pooling overtime of all the convolutional filter outputs.

6 Benchmark results and discussion

In this section, we describe our live blog summarization experiments and provide benchmark results for future researchers using our data and setup.

6.1 Experimental setup

In our experiments, we measure performance using the ROUGE metrics identified by Owczarzak et al. (2012) as strongly correlating with human evaluation methods: ROUGE-1 (R1), ROUGE-2 (R2) and ROUGE-L (RL) recall with stemming and stop words not removed. We explore two different summary lengths: 50 words, which corresponds to the average length of the human-written summary, and 100 words, which is twice the average length of the human-written summaries in order to give leeway for compensating the excessive compression ratio of the human-written live blog summaries.

Table 7 Training, validation and test split sizes for BBC and Guardian datasets

Dataset	Train	Valid	Test
BBC	610	77	75
Guardian	1350	167	166

For the supervised setup, we split the dataset into training, validation and testing consisting of 80%, 10%, and 10% of the data respectively. Table 7 illustrates the training, validation, and test split sizes used for our experiments.

We train the models to minimize the weighted negative log-likelihood over the training data D :

$$\mathcal{L} = - \sum_{s,y \in D} \sum_{i=1}^n \omega(y_i) \log p(y_i | y_{\leq i}, h), \text{ where } h = \text{enc}(s)$$

We use the stochastic gradient descent with the Adam optimizer for optimizing the objective function. $\omega(y)$ represents the weights of the labels i.e. $\omega(0) = 1$ and $\omega(1) = \frac{N_0}{N_1}$ where N_y is the number of training samples with label y . The word embeddings are initialized using the pretrained GloVe embeddings (Pennington et al. 2014) and not updated during training. The training is carried out for a maximum of 50 epochs and the best model is selected using an early stopping criterion for ROUGE-2 on the validation set. We use a learning rate of 0.0001, a dropout rate of 0.25, and bias terms of 0. The batch size is set to 32 for both BBC and the Guardian. Additionally, due to the GPU memory limitation, the number of input sentences used by the extractors is set to 250 for BBC and 200 for the Guardian.

6.2 Upper bound

For comparison, we compute two upper bounds UB-1 and UB-2. The upper bound for extractive summarization is retrieved by solving the maximum coverage of n-grams from the reference summary (Takamura and Okumura 2010; Peyrard and Eckle-Kohler 2016; Avinesh and Meyer 2017). Upper bound summary extraction is cast as an ILP problem as described in Eqs. (2–7), which is the core of the ICSI system. However, the only difference is that the concept weights are set to 1 if the concepts occur in the human-written reference summary. The concept extraction depends on N , which represents the n-gram concept type. In our work, we set $N = 1$ and $N = 2$ and compute the upper bound for ROUGE-1 (UB-1) and ROUGE-2 (UB-2) respectively.

6.3 Analysis

Table 8 shows the benchmark results of the five unsupervised summarization methods introduced in Sect. 5.1 on our live blog corpus in comparison to the

Table 8 ROUGE-1 (R1), ROUGE-2 (R2), and ROUGE-L (RL) scores of multiple unsupervised systems compared to the extractive upper bounds for ROUGE-1 (UB-1) and ROUGE-2 (UB-2) for summary lengths of 50 and 100 words

Systems	BBC			Guardian			DUC 2004								
	50 words			50 words			100 words								
	R1	R2	RL	R1	R2	RL	R1	R2	RL						
TF-IDF	0.184	0.030	0.114	0.274	0.056	0.155	0.158	0.015	0.104	0.245	0.028	0.153	0.292	0.055	0.186
LexRank	0.208	0.042	0.132	0.308	0.080	0.181	0.198	0.022	0.129	0.292	0.039	0.177	0.345	0.070	0.208
LSA	0.176	0.018	0.018	0.257	0.035	0.144	0.143	0.010	0.100	0.229	0.020	0.141	0.294	0.045	0.181
KL	0.193	0.032	0.118	0.274	0.053	0.160	0.172	0.019	0.116	0.256	0.030	0.159	0.336	0.072	0.204
ICSI	0.277	0.079	0.180	0.374	0.111	0.214	0.223	0.038	0.140	0.320	0.050	0.194	0.374	0.090	0.218
UB-1	0.439	0.184	0.250	0.622	0.272	0.301	0.367	0.085	0.207	0.536	0.119	0.269	0.492	0.183	0.265
UB-2	0.419	0.230	0.263	0.576	0.331	0.304	0.313	0.134	0.201	0.429	0.185	0.250	0.472	0.210	0.282

Bold face indicates best system performance per column and relevant upper bound

standard DUC 2004 dataset. TF-IDF and LSA consistently lag behind the other methods. The results of KL are in a mid-range for the DUC datasets, but low on our data. LexRank yields stable results, but ICSI as a state-of-the-art method for unsupervised extractive summarization consistently outperforms all other methods by a large margin. The automatic methods reach higher ROUGE scores on BBC than on Guardian data, which we attribute to the different level of abstractiveness used for these live blogs: In BBC and DUC 2004, the summaries tend to reuse verbatim phrases from the input documents, whereas the Guardian summaries often contain newly formulated sentences in the summary. This can also be observed in the upper bound, as both UB-1 and UB-2 for the Guardian data are lower than the corresponding values for BBC. The best unsupervised method ICSI is 0.15 ROUGE-1 and 0.2 ROUGE-2 lower than the upper bounds for BBC and 0.1 ROUGE-1 and 0.1 ROUGE-2 lower for the Guardian's upper bounds.

The results of the supervised approaches comparing different extractors and encoders are shown in Table 9. While ICSI is the only unsupervised approach which is able to reach one-third of the upper bound, supervised approaches can reach up to 50% of the upper bound scores for BBC. This confirms that the supervised models are able to learn importance properties of the BBC dataset. However, the supervised models perform worse than ICSI on the Guardian dataset. We presume this is caused by the constraint on the number of input sentences due to the GPU memory constraint.

Overall, there are improvements of about 0.03 ROUGE-1 and 0.02 ROUGE-2 when a CNN encoder is used for sentence representation as compared to the averaging encoder across all the supervised approaches, which differs from the observation by Kedzie et al. (2018). When analyzing different extractors, the Seq2Seq extractor performs best in the majority of the settings, closely followed by Cheng and Lapata and RNN. SummRuNNer consistently yields lower scores across all settings. Although RNN yields slightly better results on the 100 words condition of the Guardian data, Seq2Seq and Cheng and Lapata with CNN encoder yield consistently good results across both datasets.

Figure 4 shows the output of the best unsupervised system ICSI and the three best supervised systems (i.e. Chang and Lapata, RNN, and Seq2Seq with a CNN encoder). The outputs are compared to the extractive upper bound UB-2 and the reference summary for the BBC live blog on “Junior doctors’ strike updates”.¹⁴ ICSI extracts sentences with the most frequent concepts (e.g., junior doctor, strike, England), but misses to identify topic shifts in the live blog’s postings, such as the discussion of emergency cover. The best supervised approach Seq2Seq captures more diverse concepts (e.g. junior doctors, emergency cover, 24-h walkout, dispute with the government) covering a greater variety of information about the strike event and its agents and reasons.

However, the example also shows the challenges of live blog summarization, since most methods incorporate general statements to capture the reader’s attention (e.g., “stay with us as we bring you the latest updates”), which contain little factual information, but are frequently found in the postings. Many of our methods failed to

¹⁴ <https://www.bbc.com/news/live/health-35290222> (accessed January 16, 2019).

Table 9 ROUGE-1 (R1), ROUGE-2 (R2), and ROUGE-L (L) scores across supervised neural methods with all extractor and encoder (enc.) pairs compared to the extractive upper bounds for ROUGE-1 (UB-1) and ROUGE-2 (UB-2)

Extractor	Enc.	BBC						Guardian					
		50 words			100 words			50 words			100 words		
		R1	R2	RL	R1	R2	RL	R1	R2	RL	R1	R2	RL
RNN	Avg.	0.283	0.078	0.156	0.379	0.110	0.250	0.174	0.019	0.040	0.257	0.028	0.062
	CNN	0.296	0.095	0.164	0.390	0.123	0.151	0.181	0.019	0.040	0.273	0.034	0.067
Seq2Seq	Avg.	0.287	0.083	0.161	0.380	0.109	0.246	0.175	0.020	0.046	0.254	0.024	0.060
	CNN	0.296	0.093	0.162	0.400	0.130	0.261	0.184	0.023	0.047	0.269	0.031	0.063
Cheng and Lapata	Avg.	0.279	0.080	0.155	0.372	0.108	0.242	0.177	0.020	0.048	0.254	0.027	0.061
	CNN	0.305	0.105	0.174	0.383	0.121	0.249	0.181	0.020	0.048	0.270	0.030	0.064
SummaRuNNer	Avg.	0.245	0.055	0.125	0.331	0.067	0.204	0.161	0.014	0.030	0.224	0.021	0.058
	CNN	0.274	0.080	0.144	0.383	0.115	0.248	0.172	0.017	0.031	0.256	0.027	0.061
UB-1	-	0.439	0.184	0.250	0.622	0.272	0.301	0.367	0.085	0.207	0.536	0.119	0.269
UB-2	-	0.419	0.230	0.263	0.576	0.331	0.304	0.313	0.134	0.201	0.429	0.185	0.250

Bold face indicates best system performance per column and relevant upper bound

<p>Junior doctors in England are taking part in a 24-hour strike on Tuesday 12 January 2016 in dispute with government. Emergency cover only being provided after 08:00 GMT. There are 55,000 junior doctors – about a third of the workforce. Three strikes are planned – the last in February will see doctors refuse to provide emergency care.</p>	<p>This is not surprising as doctors had agreed to provide emergency care cover. There are 55,000 junior doctors in England, which is about a third of the workforce. They are taking part in a 24-hour strike in a dispute with the government over a new contract.</p>
<p>(a) Reference</p>	<p>(b) Upper bound UB-2</p>
<p>She also says the government’s action on changing contracts was a step towards privatising the NHS. Want to know more about what’s going on with the junior doctor strike in England? @twitterid thank you Noel! Here’s a bit more from Jon Stanley, a junior doctor who isn’t supporting the strike.</p>	<p>They are taking part in a 24-hour strike in a dispute with the government over a new contract. Stay with us as we bring you the latest updates, images and tweets covering the strike. Junior doctors will provide emergency cover only during the 24-hour walkout, which got under way at 08:00 GMT.</p>
<p>(c) ICSI</p>	<p>(d) Seq2Seq + CNN</p>
<p>This is our coverage of today’s industrial action by junior doctors. Junior doctors will provide emergency cover only during the 24-hour walkout, which got under way at 08:00 GMT. Stay with us as we bring you the latest updates, images and tweets covering the strike. little do with patients - it’s a middle class fight to preserve week day working - now mostly reserved for offices @twitterid @twitterid support the doctors.</p>	<p>They are taking part in a 24-hour strike in a dispute with the government over a new contract. Stay with us as we bring you the latest updates, images and tweets covering the strike. Tests, appointments and clinics are also being hit, and an estimated one in 10 non-emergency patients look like they will be affected on the day.</p>
<p>(e) Chang & Lapata + CNN</p>	<p>(f) RNN + CNN</p>

Fig. 4 System outputs on the BBC.com live blog on Junior doctors’ strike updates

detect this raising the need for methods that better take semantic aspects into account. Furthermore, none of the summaries provides information about the greater context and future outlook (i.e., the fact that three strikes are planned). Such information is very important for summarizing live blogs, since readers are typically interested into the implications of certain events or decisions. The same applies to quotes by major protagonists of an event, as they are often included in a live blog summary, but not yet particularly treated by the automatic summarization methods. The increasing use of multimedia also raises a need for multimodal approaches that are able to extract important content from images or videos and include them into a summary. For multimodal summarization, there are yet only few case studies for a few domains, such as financial reports (Ahmad et al. 2004). Among the biggest challenges is, however, the heterogeneity of the individual postings, which makes the task of live blog summarization much different to multi-document summarization of multiple news articles covering very similar information or microblog summarization of a large number of highly redundant posts. In live blogs, the same fact is typically covered only once.

7 Conclusion and future work

Automatic live blog summarization is a new task with direct applications for journalists and news readers, as journalists can easily summarize the major facts about an event and even provide instant updates as intermediate summaries while the event is ongoing. In this paper, we suggest a pipeline to collect live blogs with human-written bullet-point summaries from two major online newspapers, the BBC and the Guardian. Our pipeline can be extended to collect live blogs from other news agencies as well, including the *New York Times*, the *Washington Post* or *Der Spiegel*.

Based on this live blog reference corpus, we analyze the domain distribution and the heterogeneity of the corpus, and we provide benchmark results using state-of-the-art summarization methods. Our results show that simple off-the-shelf unsupervised summarization systems are not very effective for live blog summarization. Supervised systems, however, yield better results, particularly on our BBC data. We find the Seq2Seq extractor with a CNN encoder for sentence representations to perform best in the majority of settings. Furthermore, sentence representations based on a CNN encoder show improvements of 0.03 ROUGE-1 0.02 ROUGE-2 compared to the averaging encoder. For the Guardian data, the supervised systems showed worse results than the unsupervised ICSI system. Our results enable future research on novel approaches to live blog summarization that are able to successfully handle the large number of heterogeneous postings of a live blog.

Besides our benchmark results which allow for comparison, we provide the source code for constructing and reproducing the live blog corpus as well as the automatic summarization experiments under the permissive Apache License 2.0 from our GitHub repository <https://github.com/AIPHES/live-blog-summarization>.

Acknowledgements This work has been supported by the German Research Foundation as part of the Research Training Group “Adaptive Preparation of Information from Heterogeneous Sources” (AIPHES) under Grant No. GRK 1994/1. We also acknowledge the useful comments and suggestions of the anonymous reviewers.

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Ahmad, S., de Oliveira, P. C. F., & Ahmad, K. (2004). Summarization of multimodal information. In *Proceedings of the fourth international conference on language resources and evaluation (LREC), Lisbon, Portugal* (pp. 1049–1052). <http://www.lrec-conf.org/proceedings/lrec2004/>.
- Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., et al. (2017). Text summarization techniques: A brief survey. *International Journal of Advanced Computer Science and Applications*, 8(10), 397–405. <https://doi.org/10.14569/IJACSA.2017.081052>.
- Aone, C., Okurovski, M. E., Gorlinsky, J., & Larsen, B. (1995). A trainable summarizer with knowledge acquired from robust NLP techniques. In I. Mani & M. T. Maybury (Eds.), *Advances in automatic text summarization* (pp. 68–73). Cambridge, MA: MIT Press.
- Avinesh, P. V. S., & Meyer, C. M. (2017). Joint optimization of user-desired content in multi-document summaries by learning from user feedback. In *Proceedings of the 55th annual meeting of the association for computational linguistics (ACL), Vancouver, BC, Canada* (pp. 1353–1363). <https://doi.org/10.18653/v1/P17-1124>.
- Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the international conference on learning representations (ICLR), San Diego, CA, USA*. <https://arxiv.org/abs/1409.0473>.
- Baroni, M., & Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of the 4th international conference on language resources and evaluation (LREC), Lisbon, Portugal* (pp. 1313–1316). <http://lrec-conf.org/proceedings/lrec2004/summaries/509.htm>.
- Barzilay, R., & McKeown, K. R. (2005). Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3), 297–328. <https://doi.org/10.1162/089120105774321091>.
- Baumer, E., Elovic, E., Qin, Y., Polletta, F., & Gay, G. (2015). Testing and comparing computational approaches for identifying the language of framing in political news. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies (NAACL/HLT), Denver, CO, USA* (pp. 1472–1482). <https://aclweb.org/anthology/N15-1171>.
- Benikova, D., Mieskes, M., Meyer, C. M., & Gurevych, I. (2016). Bridging the gap between extractive and abstractive summaries: Creation and evaluation of coherent extracts from heterogeneous sources. In *Proceedings of the 26th international conference on computational linguistics (COLING), Osaka, Japan* (pp. 1039–1050). <https://aclweb.org/anthology/C16-1099>.
- Bird, S., Dale, R., Dorr, B. J., Gibson, B., Joseph, M. T., Kan, M. Y., Lee, D., Powley, B., Radev, D. R., & Tan, Y. F. (2008). The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the sixth international conference on language resources and evaluation (LREC), Marrakech, Morocco* (pp. 1755–1759). <http://lrec-conf.org/proceedings/lrec2008/summaries/445.html>.
- Birnbbaum, L., Popescu, O., & Strapparava, C. (Eds.). (2016). *Proceedings of the workshop on natural language processing meets journalism, New York, NY, USA*. <http://nlpj2016.fbk.eu/proceedings>.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022. <http://jmlr.org/papers/v3/blei03a.html>.
- Blom, J. N., & Hansen, K. R. (2015). Click bait: Forward-reference as lure in online news headlines. *Journal of Pragmatics*, 76, 87–100. <https://doi.org/10.1016/j.pragma.2014.11.010>.
- Bourgonje, P., Moreno Schneider, J., & Rehm, G. (2017). From clickbait to fake news detection: An approach based on detecting the stance of headlines to articles. In *Proceedings of the Second Workshop on Natural Language Processing meets Journalism, Copenhagen, Denmark* (pp 84–89). <https://doi.org/10.18653/v1/W17-4215>.
- Bozdag, E., & van den Hoven, J. (2015). Breaking the filter bubble: Democracy and design. *Ethics and Information Technology*, 17(4), 249–265. <https://doi.org/10.1007/s10676-015-9380-y>.
- Brandtzaeg, P. B., Lüders, M., Spangenberg, J., Rath-Wiggings, L., & Flstad, A. (2015). Emerging journalistic verification practices concerning social media. *Journalism Practice*, 10(3), 323–342. <https://doi.org/10.1080/17512786.2015.1020331>.
- Cao, Z., Chen, C., Li, W., Li, S., Wei, F., & Zhou, M. (2016). TGSUM: Build tweet guided multi-document summarization dataset. In *Proceedings of the thirtieth conference on artificial intelligence (AAAI), Phoenix, AZ, USA* (pp. 2906–2912). <https://aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11991>.

- Cao, Z., Wei, F., Dong, L., Li, S., & Zhou, M. (2015). Ranking with recursive neural networks and its application to multi-document summarization. In *Proceedings of the twenty-ninth conference on artificial intelligence (AAAI)*, Austin, TX, USA (pp. 2153–2159). <https://aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9414>.
- Cheng, J., & Lapata, M. (2016). Neural summarization by extracting sentences and words. In *Proceedings of the 54th annual meeting of the association for computational linguistics (ACL)*, Berlin, Germany. <https://aclweb.org/anthology/P16-1046>.
- Chopra, S., Auli, M., & Rush, A. M. (2016). Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies (NAACL/HLT)*, San Diego, CA, USA (pp. 93–98). <https://aclweb.org/anthology/N16-1012>.
- Clarke, I., & Grieve, J. (2017). Dimensions of abusive language on Twitter. In *Proceedings of the first workshop on abusive language online*, Vancouver, BC, Canada (pp. 1–10). <https://aclweb.org/anthology/W17-3001>.
- Conroy, J. M., & O’Leary, D. P. (2001). Text summarization via hidden markov models. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval*, New Orleans, LA, USA (pp. 406–407). <https://doi.org/10.1145/383952.384042>.
- Dang, H., & Owczarzak, K. (2008). Overview of the TAC 2008 update summarization task. In *Proceedings of the first text analysis conference (TAC)*, Gaithersburg, MD, USA (pp. 1–16). http://www.nist.gov/tac/publications/2008/additional.papers/update_summ_overview08.proceedings.pdf.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391–407.
- Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61–74. <https://aclweb.org/anthology/J93-1003>.
- Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22, 457–479. <https://www.aaai.org/Papers/JAIR/Vol22/JAIR-2214.pdf>.
- Filippova, K. (2010). Multi-sentence compression: Finding shortest paths in word graphs. In *Proceedings of the 23rd international conference on computational linguistics (COLING)*, Beijing, China (pp. 322–330). <https://aclweb.org/anthology/C10-1037>.
- Fu, L., Danescu-Niculescu-Mizil, C., & Lee, L. (2016). Tie-breaker: Using language models to quantify gender bias in sports journalism. In *Proceedings of the workshop on natural language processing meets journalism*, New York, NY, USA (pp. 1–5).
- Gatti, L., Özbal, G., Guerini, M., Stock, O., & Strapparava, C. (2016). Automatic creation of flexible catchy headlines. In *Proceedings of the workshop on natural language processing meets journalism*, New York, NY, USA (pp. 25–29).
- Giannakopoulos, G., Kubina, J., Conroy, J., Steinberger, J., Favre, B., Kabadjov, M., Kruschwitz, U., & Poesio, M. (2015). MultiLing 2015: Multilingual summarization of single and multi-documents, on-line fora, and call-center conversations. In *Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue*, Prague, Czech Republic (pp. 270–274). <https://aclweb.org/anthology/W15-4638>.
- Gillick, D., & Favre, B. (2009). A scalable global model for summarization. In *Proceedings of the workshop on integer linear programming for natural language processing*, Boulder, CO, USA (pp. 10–18). <https://aclweb.org/anthology/W09-1802>.
- Gong, Y., & Liu, X. (2001). Generic analysis. In *Proceedings of the 24th International ACM SIGIR Ltrieval*, New Orleans, LA, USA (pp. 19–25). <https://doi.org/10.1145/383952.383955>.
- Gu, J., Lu, Z., Li, H., & Li, V. O. (2016). Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th annual meeting of the association for computational linguistics*, Berlin, Germany (pp. 1631–1640). <https://aclweb.org/anthology/P16-1154>.
- Gulcehre, C., Ahn, S., Nallapati, R., Zhou, B., & Bengio, Y. (2016). Pointing the unknown words. In *Proceedings of the 54th annual meeting of the association for computational linguistics*, Berlin, Germany (pp. 140–149). <http://www.aclweb.org/anthology/P16-1014>.
- Haghighi, A., & Vanderwende, L. (2009). Exploring content models for multi-document summarization. In *Proceedings of the 2009 annual conference of the North American chapter of the association for computational linguistics: Human language technologies (NAACL/HLT)*, Boulder, CO, USA (pp. 362–370). <https://aclweb.org/anthology/N09-1041>.

- Hakkani-Tur, D., & Tur, G. (2007). Statistical sentence extraction for information distillation. In *IEEE international conference on acoustics, speech and signal processing, Honolulu, HI, USA* (Vol. IV, pp. 1–4). <https://doi.org/10.1109/ICASSP.2007.367148>.
- Hanselowski, A., PVS, A., Schiller, B., Caspelherr, F., Chaudhuri, D., Meyer, C. M., & Gurevych, I. (2018). A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th international conference on computational linguistics (COLING), Santa Fe, NM, USA* (pp. 1859–1874). <https://aclweb.org/anthology/C18-1158>.
- Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching machines to read and comprehend. In *Proceedings of the 28th international conference on neural information processing systems (NIPS), Montreal, QC, Canada* (pp. 1693–1701). <https://papers.nips.cc/paper/5945-teaching-machines-to-read-and-comprehend.pdf>.
- Hong, K., Conroy, J., Favre, B., Kulesza, A., Lin, H., & Nenkova, A. (2014). A repository of state of the art and competitive baseline summaries for generic news summarization. In *Proceedings of the ninth international conference on language resources and evaluation (LREC), Reykjavik, Iceland* (pp. 1608–1616). <http://lrec-conf.org/proceedings/lrec2014/summaries/1093.html>.
- Hong, K., & Nenkova, A. (2014). Improving the estimation of word importance for news multi-document summarization. In *Proceedings of the 14th conference of the European chapter of the association for computational linguistics (EACL), Gothenburg, Sweden* (pp. 712–721). <https://aclweb.org/anthology/E14-1075>.
- Hovy, E., & Lin, C. Y. (1999). Automated text summarization and the SUMMARIST system. In I. Mani & M. T. Maybury (Eds.), *Advances in automatic text summarization* (pp. 82–94). Cambridge: MIT Press.
- Kedzie, C., McKeown, K. R., & Daume III, H. D. (2018). Content selection in deep learning models of summarization. In *Proceedings of the 2018 conference on empirical methods in natural language processing (EMNLP), Brussels, Belgium* (pp. 1818–1828). <https://aclweb.org/anthology/D18-108>.
- Khan, A., Salim, N., & Farman, H. (2016). Clustered genetic semantic graph approach for multi-document abstractive summarization. In *International conference on intelligent systems engineering (ICISE), Islamabad, Pakistan* (pp. 63–70). <https://doi.org/10.1109/INTELSE.2016.7475163>.
- Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar* (pp. 1746–1751). <https://aclweb.org/anthology/D14-1181>.
- Knight, K., & Marcu, D. (2000). Statistics-based summarization—step one: Sentence compression. In *Proceedings of the seventeenth national conference on artificial intelligence (AAAI), Austin, TX, USA* (pp. 703–710). <https://aaai.org/Papers/AAAI/2000/AAAI00-108.pdf>.
- Ko, Y., & Seo, J. (2008). An effective sentence-extraction technique using contextual information and statistical approaches for text summarization. *Pattern Recognition Letters*, 29(9), 1366–1371. <https://doi.org/10.1016/j.patrec.2008.02.008>.
- Kolhatkar, V., & Taboada, M. (2017). Using New York times picks to identify constructive comments. In *Proceedings of the second workshop on natural language processing meets journalism, Copenhagen, Denmark* (pp. 100–105). <https://doi.org/10.18653/v1/W17-4218>.
- Kuang, S., & Davison, B. (2016). Semantic and context-aware linguistic model for bias detection. In *Proceedings of the workshop on natural language processing meets journalism, New York, NY, USA* (pp. 57–62).
- Kupiec, J., Pedersen, J., & Chen, F. (1995). A trainable document summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval, Seattle, WA, USA* (pp. 68–73). <https://doi.org/10.1145/215206.215333>.
- Li, C., Qian, X., & Liu, Y. (2013). Using supervised bigram-based ILP for extractive summarization. In *Proceedings of the 51st annual meeting of the association for computational linguistics (ACL), Sofia, Bulgaria* (pp. 1004–1013). <https://aclweb.org/anthology/N15-1145>.
- Li, L., Zhou, K., Xue, G. R., Zha, H., & Yu, Y. (2009). Enhancing diversity, coverage and balance for summarization through structure learning. In *Proceedings of the 18th international conference on world wide web (WWW), Madrid, Spain* (pp. 71–80). <https://doi.org/10.1145/1526709.1526720>.
- Li, P., Bing, L., & Lam, W. (2017). Reader-aware multi-document summarization: An enhanced model and the first dataset. In *Proceedings of the EMNLP workshop on new frontiers in summarization, Copenhagen, Denmark* (pp. 91–99). <https://aclweb.org/anthology/W17-4512>.

- Lin, J., Roegiest, A., Tan, L., McCreddie, R., Voorhees, E., & Diaz, F. (2016). Overview of the trec 2016 real-time summarization track. In *Proceedings of the twenty-fifth text retrieval conference (TREC)*, Gaithersburg, MD, USA. <http://trec.nist.gov/pubs/trec25/papers/Overview-RT.pdf>.
- Liu, F., Liu, Y., & Weng, F. (2011). Why is “SXSWS” trending? Exploring multiple text sources for twitter topic summarization. In *Proceedings of the ACL workshop on language in social media*, Portland, OR, USA (pp. 66–75). <https://www.aclweb.org/anthology/W11-0709>.
- Liu, X., Li, Y., Wei, F., & Zhou, M. (2012). Graph-based multi-tweet summarization using social signals. In *Proceedings of the 24th international conference on computational linguistics (COLING)*, Mumbai, India (pp. 1699–1714). <https://www.aclweb.org/anthology/C12-1104>.
- Lloret, E., & Palomar, M. (2013). Towards automatic tweet generation: A comparative study from the text summarization perspective in the journalism genre. *Expert Systems with Applications*, 40(16), 6624–6630. <https://doi.org/10.1016/j.eswa.2013.06.021>.
- Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2), 159–165. <https://doi.org/10.1147/rd.22.0159>.
- Mani, I., & Bloedorn, E. (1997). Multi-document summarization by graph search and matching. In *Proceedings of the fourteenth national conference on artificial intelligence and ninth conference on innovative applications of artificial intelligence (IAAI)*, Providence, RI, US (pp. 622–628). <https://aaai.org/Papers/AAAI/1997/AAAI97-097.pdf>.
- Mani, I., & Bloedorn, E. (1999). Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1–2), 35–67. <https://doi.org/10.1023/A:1009930203452>.
- McDonald, R. (2007). A study of global inference algorithms in multi-document summarization. In *Proceedings of the 29th European conference on IR research (ECIR)*, Springer, Rome, Italy (pp. 557–564). https://doi.org/10.1007/978-3-540-71496-5_51.
- McEnnis, S. (2016). Following the action: How live bloggers are reimagining the professional ideology of sports journalism. *Journalism Practice*, 10(8), 967–982. <https://doi.org/10.1080/17512786.2015.1068130>.
- Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing (EMNLP)*, Barcelona, Spain (pp. 404–411). <https://aclweb.org/anthology/W04-3252>.
- Nallapati, R., Zhai, F., & Zhou, B. (2017). SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the thirty-first conference on artificial intelligence (AAAI)*, San Francisco, CA, USA (pp. 3075–3081). <https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14636/14080>.
- Nallapati, R., Zhou, B., dos Santos, C., Gulcehre, C., & Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL conference on computational natural language learning (CoNLL)*, Berlin, Germany (pp. 280–290). <https://aclweb.org/anthology/K16-1028>.
- Nenkova, A., & Louis, A. (2008). Can you summarize this? Identifying correlates of input difficulty for multi-document summarization. In *Proceedings of the 46th annual meeting of the association for computational linguistics (ACL)*, Columbus, OH, USA (pp. 825–833). <https://aclweb.org/anthology/P08-1094>.
- Nenkova, A., & McKeown, K. R. (2012). A survey of text summarization techniques. In C. C. Aggarwal & C. Zhai (Eds.), *Mining text data* (pp. 43–76). Boston: Springer. https://doi.org/10.1007/978-1-4614-3223-4_3.
- Nenkova, A., Vanderwende, L., & McKeown, K. (2006). A compositional context sensitive multi-document summarizer: Exploring the factors that influence summarization. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval*, Seattle, WA, USA (pp. 573–580). <https://doi.org/10.1145/1148170.1148269>.
- Owczarzak, K., Conroy, J. M., Dang, H. T., & Nenkova, A. (2012). An assessment of the accuracy of automatic evaluation in summarization. In *Proceedings of workshop on evaluation metrics and system comparison for automatic summarization*, Montréal, QC, Canada (pp. 1–9). <https://aclweb.org/anthology/W12-2601>.
- Pavlopoulos, J., Malakasiotis, P., Bakagianni, J., & Androutopoulos, I. (2017). Improved abusive comment moderation with user embeddings. In *Proceedings of the second workshop on natural language processing meets journalism*, Copenhagen, Denmark (pp. 51–55). <https://doi.org/10.18653/v1/W17-4209>.

- Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, Doha, Qatar (pp. 1532–1543). <https://aclweb.org/anthology/D14-1162>.
- Peyrard, M., & Eckle-Kohler, J. (2016). Optimizing an approximation of ROUGE—a problem-reduction approach to extractive multi-document summarization. In *Proceedings of the 54th annual meeting of the association for computational linguistics (ACL)*, Berlin, Germany (pp. 1825–1836). <https://aclweb.org/anthology/P16-1172>.
- Peyrard, M., & Eckle-Kohler, J. (2017). Supervised learning of automatic pyramid for optimization-based multi-document summarization. In *Proceedings of the 55th annual meeting of the association for computational linguistics (ACL)*, Vancouver, BC, Canada (Vol. 1, pp. 1084–1094). <https://doi.org/10.18653/v1/P17-1100>.
- Popescu, O., & Strapparava, C. (Eds.). (2017). *Proceedings of the second workshop on natural language processing meets journalism, Copenhagen, Denmark*. <https://aclweb.org/anthology/W17-4200>.
- Popescu, O., & Strapparava, C. (Eds.). (2018). *Proceedings of the third workshop on natural language processing meets journalism, Miyazaki, Japan*. <http://lrec-conf.org/workshops/lrec2018/W13/>.
- Potash, P., Romanov, A., Gronas, M., Rumshisky, A., & Gronas, M. (2017). Tracking bias in news sources using social media: The Russia-Ukraine maidan crisis of 2013–2014. In *Proceedings of the second workshop on natural language processing meets journalism, Copenhagen, Denmark* (pp. 13–18). <https://doi.org/10.18653/v1/W17-4203>.
- Radev, D. R., Hovy, E., & McKeown, K. R. (2002). Introduction to the special issue on summarization. *Computational Linguistics*, 28(4), 399–408. <https://doi.org/10.1162/089120102762671927>.
- Radev, D. R., Jing, H., & Budzikowska, M. (2000). Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies. In *Proceedings of the NAACL-ANLP workshop on automatic summarization, Seattle, Washington* (pp. 21–30). <https://aclweb.org/anthology/W00-0403>.
- Rush, A. M., Chopra, S., & Weston, J. (2015). A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 conference on empirical methods in natural language processing (EMNLP)*, Lisbon, Portugal (pp. 379–389). <https://doi.org/10.18653/v1/D15-1044>.
- Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media, Valencia, Spain* (pp. 1–10). <https://aclweb.org/anthology/W17-1101>.
- See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th annual meeting of the association for computational linguistics (ACL)*, Vancouver, BC, Canada (pp. 1073–1083). <https://doi.org/10.18653/v1/P17-1099>.
- Sharifi, B., Hutton, M. A., & Kalita, J. K. (2010). Experiments in microblog summarization. In *Proceedings of the 2010 IEEE second international conference on social computing, Minneapolis, MN, USA* (pp. 49–56). <https://doi.org/10.1109/SocialCom.2010.17>.
- Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11–21.
- Steinberger, J., & Jezek, K. (2004). Using latent semantic analysis in text summarization and summary evaluation. In *Proceedings of the 7th international conference on information systems implementation and modelling (ISIM)*, Rožnov pod Radhoštěm, Czech Republic (pp. 93–100).
- Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of the 27th international conference on neural information processing systems (NIPS)*, Montreal, QC, Canada (pp. 3104–3112). <https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf>.
- Szymanski, T., Orellana-Rodriguez, C., & Keane, M. (2016). Helping news editors write better headlines: A recommender to improve the keyword contents & shareability of news headlines. In *Proceedings of the workshop on natural language processing meets journalism, New York, NY, USA* (pp. 30–34).
- Takamura, H., & Okumura, M. (2010). Learning to generate summary as structured output. In *Proceedings of the 19th ACM international conference on information and knowledge management (CIKM)*, Toronto, QC, Canada (pp. 1437–1440). <https://doi.org/10.1145/1871437.1871641>.
- Tauchmann, C., Arnold, T., Hanselowski, A., Meyer, C. M., & Mieskes, M. (2018). Beyond generic summarization: A multi-faceted hierarchical summarization corpus of large heterogeneous data. In *Proceedings of the 11th international conference on language resources and evaluation (LREC)*, Miyazaki, Japan (pp. 3184–3191). <http://lrec-conf.org/proceedings/lrec2018/summaries/252.html>.

- Thorne, J., Chen, M., Myriantous, G., Pu, J., Wang, X., & Vlachos, A. (2017). Fake news stance detection using stacked ensemble of classifiers. In *Proceedings of the second workshop on natural language processing meets journalism, Copenhagen, Denmark* (pp. 80–83). <https://doi.org/10.18653/v1/W17-4214>.
- Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., & Mittal, A. (Eds.). (2018). *Proceedings of the first workshop on fact extraction and verification (FEVER), Brussels, Belgium*. <https://aclweb.org/anthology/W18-5500>.
- Thorsen, E. (2013). Live blogging and social media curation: Challenges and opportunities for journalism. In Fowler-Watt, K., & Allan, S. (Eds.). *Journalism: new challenges, Poole: Centre for Journalism & Communication Research* (Chap. 8, pp. 123–145). Poole: Bournemouth University. <http://eprints.bournemouth.ac.uk/20926/>.
- Thorsen, E., & Jackson, D. (2018). Seven characteristics defining online news formats: Towards a typology of online news and live blogs. *Digital Journalism*, 6(7), 847–868. <https://doi.org/10.1080/21670811.2018.1468722>.
- Thurman, N., & Newman, N. (2014). The future of breaking news online? A study of live blogs through surveys of their consumption, and of readers' attitudes and participation. *Journalism Studies*, 15(5), 655–667. <https://doi.org/10.1080/1461670X.2014.882080>.
- Thurman, N., & Schapals, A. K. (2017). Live blogs, sources, and objectivity: The contradictions of real-time online reporting. In *The routledge companion to digital journalism studies* (pp. 283–292). London/New York: Routledge.
- Thurman, N., & Walters, A. (2013). Live blogging—digital journalism's pivotal platform? A case study of the production, consumption, and form of Live Blogs at Guardian.co.uk *Digital Journalism*, 1(1), 82–101. <https://doi.org/10.1080/21670811.2012.714935>.
- Wilczek, B., & Blangetti, C. (2018). Live blogging about terrorist attacks: The effects of competition and editorial strategy. *Digital Journalism*, 6(3), 344–368. <https://doi.org/10.1080/21670811.2017.1359644>.
- Wong, K. F., Wu, M., & Li, W. (2008). Extractive summarization using supervised and semi-supervised learning. In *Proceedings of the 22nd international conference on computational linguistics (COLING), Manchester, UK* (Vol. 1, pp. 985–992). <https://aclweb.org/anthology/C08-1124>.
- Yao, J.G., Wan, X., & Xiao, J. (2017). Recent advances in document summarization. *Knowledge and Information Systems*, 53(2), 297–336. <https://doi.org/10.1007/s10115-017-1042-4>.
- Yin, W., & Pei, Y. (2015). Optimizing sentence modeling and selection for document summarization. In *Proceedings of the twenty-fourth international joint conference on artificial intelligence (IJCAI), Buenos Aires, Argentina* (pp. 1383–1389). <https://www.ijcai.org/Proceedings/15/Papers/199.pdf>.
- Zopf, M. (2018). auto-hMDS: Automatic construction of a large heterogeneous multilingual multi-document summarization corpus. In *Proceedings of the 11th international conference on language resources and evaluation (LREC), Miyazaki, Japan* (pp. 3228–3233). <http://lrec-conf.org/proceedings/lrec2018/summaries/1018.html>.
- Zopf, M., Peyrard, M., & Eckle-Kohler, J. (2016). The next step for multi-document summarization: A heterogeneous multi-genre corpus built with a novel construction approach. In *Proceedings of the 26th international conference on computational linguistics (COLING), Osaka, Japan* (pp. 1535–1545). <https://aclweb.org/anthology/C16-1145>.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.